Do Surrogate Endpoints Better Correlate with Overall Survival in Studies That Did Not Allow for Crossover or Reported Balanced Postprogression Treatments?

Hashim, Mahmoud; Pfeiffer, Boris M; Bartsch, Robert; Postma, Maarten; Heeg, Bart

Published in:
Value in Health

DOI:
10.1016/j.jval.2017.07.011

Publication date:
2018

VALUE IN HEALTH 21 (2018) 9–17

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Do Surrogate Endpoints Better Correlate with Overall Survival in Studies That Did Not Allow for Crossover or Reported Balanced Postprogression Treatments? An Application in Advanced Non–Small Cell Lung Cancer

Mahmoud Hashim, MPH[1,*], *Boris M. Pfeiffer, PhD*[2], *Robert Bartsch, MSc*[1], *Maarten Postma, PhD*[3], *Bart Heeg, PhD*[1]

[1]*Ingress-health, Rotterdam, The Netherlands;* [2]*Merck KgaA, Darmstadt, Germany;* [3]*University of Groningen, Groningen, The Netherlands*

ABSTRACT

Background: In previous studies, correlation between overall survival (OS) and surrogate endpoints like objective response rate (ORR) or progression-free survival (PFS) in advanced non-small cell lung cancer (NSCLC) was poor. This can be biased by crossover and postprogression treatments. **Objectives:** To evaluate the relationship between these two surrogate endpoints and OS in advanced NSCLC studies that did not allow for crossover or reported balanced post-progression treatments. **Methods:** A systematic review in patients with advanced NSCLC receiving second- and further-line therapy was performed. The relationship between the absolute difference in ORR or median PFS (mPFS) and the absolute difference in median OS (mOS) was assessed using the correlation coefficient (*R*) and weighted regression models. The analysis was repeated in predefined data cuts based on crossover and balance of postprogression treatments. When the upper limit of *R*'s 95% confidence interval (CI) was more than 0.7, the surrogate threshold effect (STE) was estimated. **Results:** In total, 146 randomized clinical trials (43,061 patients) were included. The mean ORR, mPFS, and mOS were 12.2% ± 11.2%, 3.2 ± 1.3 months, and 9.6 ± 4.1 months, respectively. The correlation coefficients of ORR and mPFS were 0.181 (95% CI 0.016–0.337) and 0.254 (95% CI 0.074–0.418), respectively, with mOS. Nevertheless, in trials that did not allow crossover and reported balanced postprogression treatments, the correlation coefficients of ORR and mPFS were 0.528 (95% CI 0.081–0.798) and 0.778 (95% CI 0.475–0.916), respectively, with mOS. On the basis of STE estimation, in trials showing significant treatment effect size of 41.0% or more ORR or 4.15 or more mPFS months, OS benefit can be expected with sufficient certainty. **Conclusions:** Crossover and postprogression treatments may bias the relationship between surrogate endpoints and OS. Presented STE calculation can be used to interpret treatment effect on either ORR or PFS when used as primary endpoints.

Keywords: crossover, non–small cell lung cancer, overall survival, surrogate endpoints validation.

Introduction

Overall survival (OS) is the criterion standard endpoint in cancer trials and is used to establish clinical benefit in support of regulatory and reimbursement applications [1–4]. Nevertheless, trials using OS as a primary endpoint need substantial sample sizes and extensive follow-up. In addition, the effects of crossover or unbalanced postprogression treatments may introduce bias or underestimate the treatment effect on OS [5,6]. An alternative surrogate endpoint for OS is progression-free survival (PFS). Regulatory agencies endorse PFS as a relevant endpoint in cancer trials [1,2,7]. In contrast to OS, PFS is not sensitive to postprogression treatments and has the advantage of assessing the duration of tumor response [5]. Objective response rate (ORR) is another potential surrogate endpoint. Compared with PFS, ORR does not assess response duration. The use of PFS and ORR as surrogate endpoints for OS would require that they be validated for this use [8]. Nevertheless, uncertainties regarding their association with OS and the potential for bias due to subjectivity in the assessment of ORR and PFS limit their use [7].

To our knowledge, only the Institute for Quality and Efficiency in Health Care (IQWiG) has issued a guidance document for surrogate endpoint validation in oncology [4]. The IQWiG recommends a stringent definition of surrogacy on the basis of the correlation coefficient (*R*). IQWiG states that if the lower limit of the 95% confidence interval (CI) of *R* is 0.85 or higher, validity of

* *Address correspondence to:* Mahmoud Hashim, Ingress-health, Health Economics and Real-World Evidence, Hofplein 20, Rotterdam 3032 AC, The Netherlands.

E-mail: mahmoud.hashim@ingress-health.com.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2017.07.011

the surrogate is suggested, but that the surrogate is not valid if the upper limit of the 95% CI is 0.7 or less [4]. Otherwise, the validity of the surrogate remains unclear; in this situation, IQWiG recommends estimating the surrogate threshold effect (STE) [4,9]. STE is defined as the minimum treatment effect on the surrogate necessary to predict a statistically significant nonzero effect on the true endpoint [9]. STE can be used to interpret the treatment effect on the surrogate endpoint.

A few studies in non–small cell lung cancer (NSCLC) have investigated the surrogacy of ORR or PFS to OS at the trial level [6,10–13]. These studies reported low correlations between PFS or ORR and OS. None of them included a stratified analysis based on the exclusion of studies allowing crossover or reporting unbalanced postprogression treatments. Stratifying studies on the basis of crossover has been done in other tumor types [14–16]. Delea et al. [15] assessed the surrogacy of PFS to OS in metastatic renal cell carcinoma trials. The correlation coefficient was greater in studies that did not allow/require crossover versus those that did allow/require crossover: correlation coefficients were estimated to be 0.50 and 0.28, respectively. Similarly, and to a less extent, greater correlation coefficients were observed in endpoint validation studies for metastatic melanoma and metastatic colorectal cancer after the removal of studies that did not allow/require crossover [14,16]. Hence, investigating the effect of crossover and postprogression treatments on the surrogacy of ORR or PFS to OS in NSCLC is warranted.

This study aimed to evaluate ORR and PFS as surrogate endpoints for OS in trials involving patients with advanced NSCLC receiving second- and further-line therapy. Then, the impact of crossover and unbalanced postprogression treatments on surrogacy was assessed.

Methods

Systematic Literature Review

The systematic literature review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement [17]. Two different bibliographic databases, PubMed and Embase, were used to identify published randomized clinical trials involving patients with stage IIIB/IV NSCLC receiving second- and further-line therapy. The search was conducted on July 28, 2016; no limitation on publication date was imposed.

A detailed search strategy (search syntax and eligibility criteria) is presented in Appendix Table 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011. One investigator reviewed the titles/abstracts of retrieved articles sequentially using the predefined eligibility criteria (see Appendix Table 1 in Supplemental Materials). Subsequently, two investigators reviewed the full text of any article that appeared to meet the eligibility criteria; disagreement was resolved by consulting with a third investigator. References in publications reviewed at the full-text stage were evaluated to identify further relevant trials.

Upon agreement on the final list of included trials, one investigator extracted data from the included trials into a predefined Microsoft Excel template. Subsequently, another investigator validated the extracted data by re-extracting them. The following data were extracted: trial identification items (e.g., PubMed identifier, first author, year, trial phase, registration identifier, and trial acronym), interventions and target population, basic patient and disease characteristics (e.g., age, sex, performance status, disease stage, histology, metastasis, and number of previous lines of therapy), additional information (e.g., use of biomarkers and crossover), and data needed for endpoint validation (number of patients in each treatment arm, ORR, PFS, and OS). Risk of bias in individual studies was assessed using the Jadad scale [18].

Assessment of Publication Bias

The risk of bias across studies was assessed using funnel plots. In this study, trial size as a measure of precision was plotted on the y-axis, and treatment effect (absolute difference) on ORR, PFS, and OS was plotted on the x-axis. In the absence of publication bias, the plot should resemble a symmetrical inverted funnel [19].

Statistical Analysis

Primary analysis

The relationship between the absolute difference in ORR and median PFS (mPFS) and the absolute difference in median OS (mOS) was assessed using the correlation coefficient (R) and weighted linear regression models. A weighted linear regression model was fitted for the following two analyses: treatment effect on ORR, with the absolute difference in ORR (%) as an independent variable (predictor) and the treatment effect on OS (absolute difference in mOS in months) as a dependent variable; and treatment effect on PFS, with the absolute difference in mPFS (months) as an independent variable (predictor) and the treatment effect on OS (absolute difference in mOS in months) as a dependent variable. Analyses were weighted by trial size, as in previous endpoint validation studies [10,13,15,20–22].

Analyses were repeated using the absolute difference in ORR (%) or PFS hazard ratio (HR) and OS-HR because HRs might capture treatment effects not captured by median survival times. We carried out log transformation of HR. Log transformation can be used to make right-skewed distributions less skewed. Treatment effect on ORR is usually reported as the absolute difference in ORR (%). For that reason and for the ease of interpretation, we used it in both analyses with OS (mOS and OS-HR). Residual versus predicted plots were inspected and diagnostic tests for normality and heteroscedasticity (nonconstant error variance) were carried out to assess consistency with the assumptions of linear regression.

First, the analysis was conducted for all trials. Trials that had allowed crossover or in which postprogression treatments were unbalanced could underestimate OS benefit and subsequently bias surrogacy evaluation. Typically, phase III trials are adequately powered for endpoints such as PFS and OS, whereas phase II trials tend to be smaller and powered for safety endpoints or ORR. Thus, phase III trials might provide more information regarding the treatment effect on these endpoints. Therefore, second, on the basis of reported postprogression treatments, we examined trial-level surrogacy in all phase III trials (data cut A), in phase III trials excluding those with per-protocol crossover (data cut B), in phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C), and in phase III trials excluding those with crossover, unbalanced postprogression treatments, or no information with regard to postprogression treatments (data cut D).

Trials that reported both the independent (the surrogate endpoint) and the dependent (the true endpoint) variables in both treatment arms were included in the analyses. For trials that included more than two treatment arms, the experimental arm was compared with a randomly chosen control arm within the same study to avoid analysis of correlated data, that is, including a treatment arm twice in the analysis. For trials that reported response in the evaluable population rather than in the intention-to-treat population, the denominator was adjusted to indicate the intention-to-treat population.

Assessing surrogacy and STE estimation

In cases in which the validity of the surrogate endpoint is deemed to be “unclear” following IQWiG guidelines [4], STE estimation is recommended to interpret treatment effect on the

Table 1 – Basic population characteristics in all included trials and prespecified data cuts based on reported postprogression therapies.

Characteristic	All trials (n = 146)		Data cut A (n = 59)		Data cut B (n = 54)		Data cut C (n = 38)		Data cut D (n = 18)	
	Valid (n)	Mean ± SD	Valid (n)	Mean ± SD	Valid (n)	Mean ± SD	Valid (n)	Mean ± SD	Valid (n)	Mean ± SD
Age (y), median*	288	61.3 ± 3.2	116	61.2 ± 3.3	106	61.2 ± 2.9	76	61.7 ± 2.8	36	61.9 ± 3.6
Male (%)*	292	63.7 ± 13.4	118	64.9 ± 13.5	108	65.5 ± 11.5	76	66.1 ± 11.1	36	60.8 ± 7.4
ECOG 0 or 1 (%)*	278	90.2 ± 12.2	116	91.2 ± 8	106	91.4 ± 8.2	74	90 ± 8.7	36	91.7 ± 8
Adenocarcinoma (%)*	238	63.3 ± 18.6	96	65.2 ± 18	90	63.7 ± 17.2	62	62.3 ± 16.7	32	71.2 ± 15
ORR (%)*	290	12.2 ± 11.2	118	12.4 ± 12.7	108	11.4 ± 11.3	76	11.3 ± 12.3	36	13.4 ± 14.5
PFS (mo)*	236	3.2 ± 1.3	100	3.2 ± 1.4	90	3.1 ± 1.3	60	3.1 ± 1.3	34	3.3 ± 1.4
OS (mo)*	282	9.6 ± 4.1	118	10 ± 4.2	108	9.6 ± 3.7	76	9.1 ± 3.8	36	10.4 ± 4.3
Jadad scale†	146	2.7 ± 1.0	59	3 ± 1.1	54	3.1 ± 1.1	38	3.1 ± 1.1	18	3.7 ± 1.3
Sample size†	146	294.9 ± 321.2	59	548.3 ± 382.3	54	567.7 ± 391.8	38	594.1 ± 439.9	18	741.6 ± 431.3

Note. Data cuts based on reported postprogression treatments: data cut A, phase III trials; data cut B, phase III trials excluding those with per-protocol crossover; data cut C, phase III trials excluding those with both per-protocol and off-protocol crossover; and data cut D, phase III trials excluding those with crossover, unbalanced postprogression treatments or insufficient information.

ECOG, Eastern Cooperative Oncology Group, performance status; n, number of observations; ORR, objective response rate; OS, overall survival; PFS, progression-free survival.

* Valid number of treatment arms with reported observation.

† Valid number of trials with reported observation.

surrogate endpoint. STE is the minimum treatment effect on the surrogate necessary to predict a statistically significant nonzero effect on the true endpoint [9]. The STE calculation allows threshold values for the decision as to whether an observed effect on the surrogate would predict (with sufficient certainty) an effect on the endpoint of interest to be specified [9]. To draw such a conclusion, the lower confidence limit of the treatment effect on the surrogate must be larger than the STE. To calculate the STE, the regression line was plotted using the weighted linear regression equation. Then, 95% prediction intervals were plotted. The value on the x-axis, the treatment effect on the surrogate, at which the lower limit of the prediction interval (upper limit in the case of relative treatment effect) meets a point corresponding to 0 on the y-axis (zero effect on the true endpoint) is the STE [9]. With stronger correlation between the surrogate endpoint and the hard endpoint, it is easier to reach the STE, for example, lower incremental mPFS months or closer to 1 PFS-HR.

Additional analyses

In addition to the likely bias due to the existence of crossover and/or unbalanced postprogression treatments, other trial or patient's characteristics may bias the quantitative relationship between surrogate endpoints and OS. Thus, first we fitted a multivariate weighted linear regression model. Such analysis would investigate whether the likely bias caused by crossover and/or unbalanced postprogression treatments still holds after adjustment for other available variables. The analysis was run only for the absolute difference in mPFS in phase III trials. The initial list of candidate independent (predictor) variables, in addition to ΔPFS, included median age, male (%), Eastern Cooperative Oncology Group (performance status) 0 or 1 (%), adenocarcinoma (%), endothelial growth factor receptor tyrosine kinase inhibitor (EGFR-TKI) treatment (dummy variable), OS primary endpoint (dummy variable), publication year, assessed Jadad scale, and data cut D (dummy variable). Crossover and postprogression treatments happen after progression. Therefore, the simultaneous influence of ΔPFS and data cut D on OS is not additive. This justifies adding an interaction term between them in the regression model.

Second, a logistic regression model was fitted with data cut D (yes = 1; no = 0) as the dependent variable. The same trial and patient's characteristics were considered as independent variables. This analysis should give more insight on the differences between studies included and excluded in the data cut D.

All analyses were carried out using the statistical software package R version 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria) and using Package "Surrogate" version 0.1-67 [23].

Results

Systematic Literature Review

Of 6274 potentially relevant publications identified, 299 hits qualified for full-text screening. After the full-text screening, 146 trials (43,061 patients) fulfilled the eligibility criteria and were

Table 2 – Association between treatment effect on ORR and PFS with OS.

Independent variable	Dependent variable	Subgroup	No. of trials	No. of patients	Correlation coefficient (R)	95% CI		STE
						Lower limit	Upper limit	
ΔORR (%)	ΔOS, median (mo)	All trials	140	41,725	0.181	0.016	0.337	NA
		Data cut A	59	32,348	0.131	0.000	0.375	NA
		Data cut B	54	30,654	0.361	0.103	0.573	NA
		Data cut C	38	22,574	0.445	0.146	0.669	NA
		Data cut D	18	13,349	0.528	0.081	0.798	41.01*
	OS-HR	All trials	76	30,570	0.172	0.000	0.383	NA
		Data cut A	44	26,549	0.374	0.086	0.604	NA
		Data cut B	41	25,534	0.399	0.104	0.629	NA
		Data cut C	27	18,854	0.521	0.175	0.752	54.86*
		Data cut D	17	13,194	0.164	0.000	0.597	NA
PFS, median (mo)	OS, median (mo)	All trials	114	35,729	0.254	0.074	0.418	NA
		Data cut A	50	27,579	0.260	0.000	0.502	NA
		Data cut B	45	25,885	0.438	0.166	0.649	NA
		Data cut C	30	18,634	0.741	0.520	0.869	3.74†
		Data cut D	17	13,194	0.778	0.475	0.916	4.15†
PFS-HR	OS-HR	All trials	73	29,907	0.402	0.190	0.579	NA
		Data cut A	42	25,386	0.463	0.185	0.672	NA
		Data cut B	39	24,371	0.461	0.170	0.678	NA
		Data cut C	25	17,691	0.694	0.412	0.855	0.24‡
		Data cut D	17	13,194	0.698	0.326	0.882	0.22‡

Note. Data cuts based on reported postprogression treatments: data cut A, all phase III trials regardless of postprogression treatments; data cut B, phase III trials excluding those with per-protocol crossover; data cut C, phase III trials excluding those with both per-protocol and off-protocol crossover; and data cut D, phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information. Δ, absolute difference; CI, confidence interval; HR, hazards ratio; NA, not available; ORR, objective response rate; OS, overall survival; PFS, progression-free survival; STE, surrogate threshold effect.

* ΔORR (%).

† ΔPFS median months.

‡ PFS-HR.

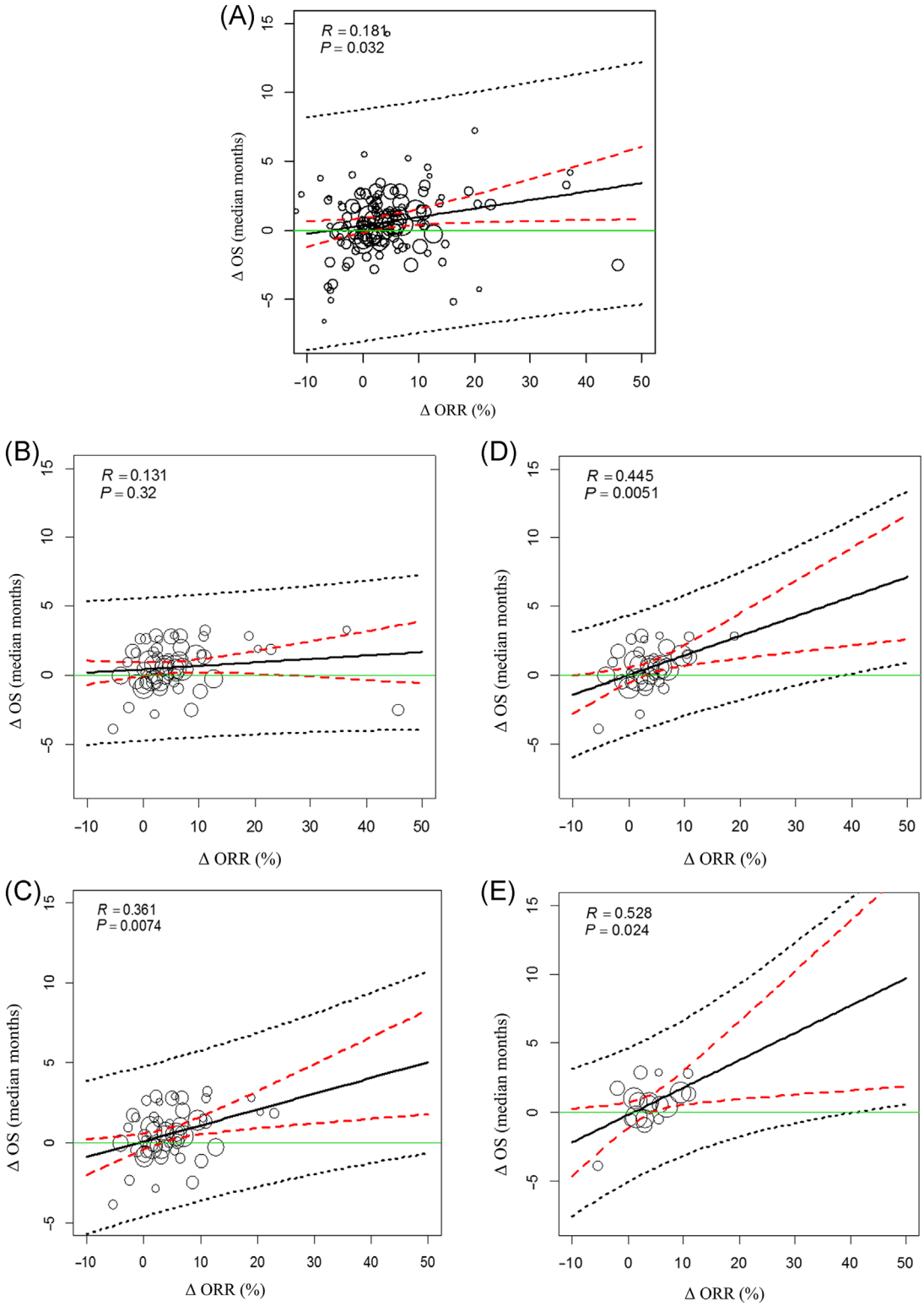

Fig. 1 – Relationship between ΔORR (*x*-axis) and ΔOS (*y*-axis): (A) primary analysis (all trials); (B) phase III trials (data cut A); (C) phase III trials excluding those with per-protocol crossover (data cut B); (D) phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C); and (E) phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information (data cut D). The solid line is the regression line. Red dashed lines are the upper and lower limits of the 95% confidence interval. Black dashed lines are the upper and lower bands of the 95% prediction intervals. Circle size is proportionate to trial size. ORR, objective response rate; OS, overall survival.

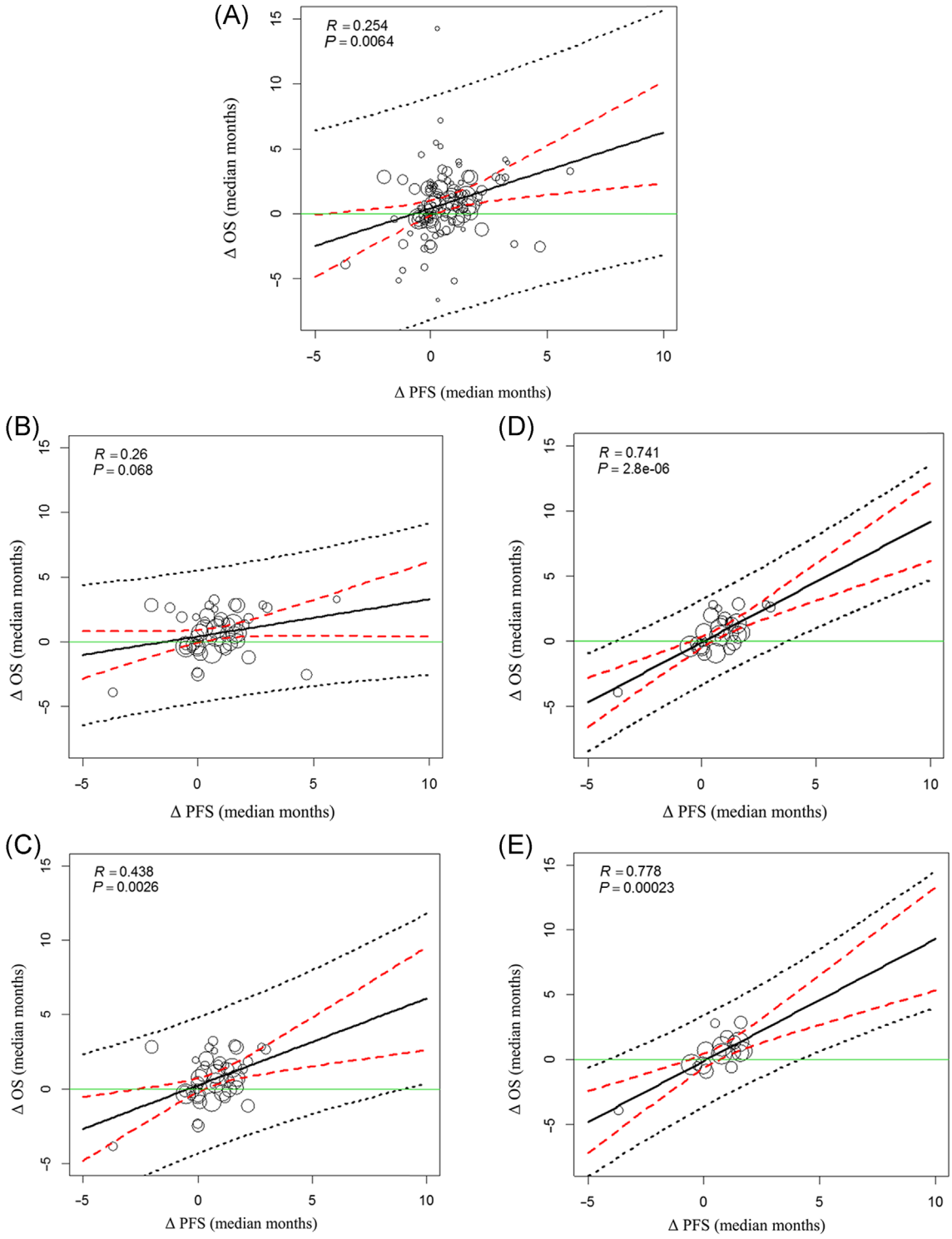

Fig. 2 – Relationship between ΔPFS (*x*-axis) and ΔOS (*y*-axis): (A) primary analysis (all trials); (B) phase III trials (data cut A); (C) phase III trials excluding those with per-protocol crossover (data cut B); (D) phase III trials excluding those with both per-protocol and off-protocol crossover (data cut C); and (E) phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information (data cut D). The solid line is the regression line. Red dashed lines are the upper and lower limits of the 95% confidence interval. Black dashed lines are the upper and lower limits of the 95% prediction intervals. Circle size is proportionate to trial size. OS, overall survival; PFS, progression-free survival.

Table 3 – Relationship between ΔPFS and ΔOS in phase III trials adjusted for patient and trial characteristics (weighted multivariate linear regression model).

Variable	Coefficient	P value	95% CI	
			Lower limit	Upper limit
ΔPFS*				
Data cut D (yes)	1.129	0.001	0.531	1.727
Data cut D (no)	0.379	0.071	−0.035	0.794
Age (y), median	0.170	0.018	0.031	0.308
ECOG 0 or 1 (%)	1.236	0.654	−4.342	6.814
Adenocarcinoma (%)†	−1.599	0.287	−4.612	1.414
EGFR treatment	−0.242	0.546	−1.051	0.567
OS primary end point	−0.240	0.570	−1.095	0.615
Publication year	−0.127	0.219	−0.334	0.079
Jadad scale	−0.227	0.273	−0.641	0.188
Constant	246.523	0.230	−164.511	657.556

Note. Data cut D is defined as phase III trials excluding those with crossover, unbalanced postprogression treatments, or insufficient information.

CI, confidence interval; ECOG, Eastern Cooperative Oncology Group, performance status; EGFR, epidermal growth factor receptor; OS, overall survival; PFS, progression-free survival.

* Interaction term between PFS and data cut D is added to the model.

† High correlation between "Male" and "Adenocarcinoma" was observed ($r = -0.808$). On the basis of collinearity diagnostics, high variance inflation factor (>5) and low tolerance (<0.2) values were observed. Therefore, we omitted the "Male" variable from the regression model.

included in the primary analysis (see Appendix Figure 1 and Appendix Tables 2 and 3 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011). Table 1 presents basic population characteristics in all included trials and in data cuts defined on the basis of postprogression treatments. Among all phase III trials (n = 59), 5 and 16 studies reported per-protocol and off-protocol crossover, respectively. These studies were excluded from data cuts A and B, respectively. Seven trials reported unbalanced postprogression treatments and 13 trials failed to report any information with regard to crossover or postprogression treatments. These studies were later excluded from data cut C.

In all treatment arms, combination therapy was the most frequent intervention (74 treatment arms) and docetaxel was the most frequent monotherapy intervention (57 treatment arms), followed by EGFR-TKIs (erlotinib/gefitinib/afatinib, 45 treatment arms) and pemetrexed (25 treatment arms). Thirty-four trials recruited exclusively Asian patients (2874 patients). A total of 87 trials involving 10,713 patients were phase II trials, whereas 59 trials involving 32,348 patients were phase III trials. The Jadad scale for individual trials was generally low (see Table 2 in Supplemental Materials); this is because most of the trials were open-label trials and sufficient information about randomization methods was not reported.

A visual examination of the funnel plots (see Appendix Figures 2–4 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011) shows that the risk of publication bias can be unlikely.

Primary Analysis

Analysis of ORR as a surrogate for OS

One hundred forty trials (41,725 patients) reported both ORR and mOS in both treatment arms. The correlation coefficient between ΔORR and ΔOS was 0.181 (95% CI 0.016–0.337) (Table 2; Fig. 1). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), the association between ΔORR and ΔOS becomes stronger: data cut A, $R = 0.131$ (95% CI 0.000–0.375); data cut B, $R = 0{\cdot}361$ (95% CI 0.103–0.573); data cut C, $R = 0.445$ (95% CI 0.146–0.669); and data cut D, $R = 0.528$ (95% CI 0.081–0.798). In data cut D, the upper limit of R's 95% CI is more than 0.7; therefore, STE was estimated to be 41% in this data cut (see Fig. 1).

Seventy-six trials (30,570 patients) reported both ORR and OS-HR in both treatment arms. The correlation coefficient between ΔORR and log(OS-HR) was 0.172 (95% CI 0.000–0.383). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, and C), association between ΔORR and log(OS-HR) becomes stronger: data cut A, $R = 0.374$ (95% CI 0.086–0.604); data cut B, $R = 0.399$ (95% CI 0.104–0.629); and data cut C, $R = 0.521$ (95% CI 0.175–0.752). In data cut C, the upper limit of R's 95% CI is more than 0.7; therefore, STE was estimated to be 55% in this data cut. This association did not achieve statistical significance in data cut D (see Appendix Figure 5 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011).

Analysis of PFS as a surrogate for OS

One hundred fourteen trials (35,729 patients) reported both mPFS and mOS in both treatment arms. The correlation coefficient between ΔPFS and ΔOS was 0.254 (95% CI 0.074–0.418) (Table 2; Fig. 2). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), association between ΔPFS and ΔOS becomes stronger: data cut A, $R = 0.260$ (95% CI 0.000–0.502); data cut B, $R = 0.438$ (95% CI 0.166–0.649); data cut C, $R = 0.741$ (95% CI 0.520–0.869); and data cut D, $R = 0.778$ (95% CI 0.475–0.916). In data cuts C and D, the upper limit of R's 95% CI is more than 0.7; therefore, STE was estimated to be 3.7 and 4.2 incremental mPFS months, respectively, in these data cuts (see Fig. 2).

Seventy-three trials (29,907 patients) reported both PFS-HR and OS-HR. The correlation coefficient between log(PFS-HR) and log(OS-HR) was 0.402 (95% CI 0.190–0.579). In further stepwise analyses based on reported postprogression treatments in phase III trials (data cuts A, B, C, and D), association between log(PFS-HR) and log(OS-HR) becomes stronger: data cut A, $R = 0.463$ (95% CI 0.185–0.672); data cut B, $R = 0.461$ (95% CI 0.170–0.678); data cut C, $R = 0.694$ (95% CI 0.412–0.855); and data cut D, $R = 0.698$ (95% CI 0.326–0.882). In data cuts C and D, the upper limit of R's 95% CI is more than 0.7; therefore, STE was estimated to be 0.24 and 0.22

PFS-HR, respectively, in these data cuts (see Appendix Figure 6 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011).

Additional Analysis

High correlation between "Male" and "Adenocarcinoma" was observed (R =−0.808) (see Appendix Table 4 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011). In addition, after running collinearity diagnostics, high variance inflation factor (>5) and low tolerance (<0.2) values were observed for both "Male" and "Adenocarcinoma." Therefore, we omitted the "Male" variable from both multivariate regression models. Table 3 presents the full multivariate linear regression model. In data cut D, an additional 1 month of ΔPFS should translate into 1.13 ΔOS months (95% CI 0.531–1.727), after adjustment for all other variables. Appendix Table 5 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.07.011 shows the results from the logistic regression model, in which data cut D is the dependent variable. Studies with no crossover and reported balanced postprogression treatments seem to be of higher quality (higher Jadad scale) and less likely to have EGFR-TKI as investigated treatment.

Discussion

This study aimed to evaluate ORR and PFS as surrogate endpoints for OS in trials involving patients with advanced NSCLC receiving second- and further-line therapy. The impact of crossover and unbalanced postprogression treatments on surrogacy was assessed. Our findings show that crossover (per- and off-protocol) and unbalanced postprogression treatments may bias the association between the surrogate endpoints such as ORR or PFS and OS. When all trials were included in the analysis, the correlation coefficients of ORR and mPFS were 0.181 (95% CI 0.016–0.337) and 0.254 (95% CI 0.074–0.418), respectively, with OS. According to the IQWiG, this suggested that ORR and PFS are not valid surrogate endpoints for OS [4]. Different results are seen in analyses in which we included trials explicitly reporting balanced postprogression treatments and excluded trials with either per-protocol or off-protocol crossover, unbalanced postprogression treatments, or no information (data cut D). Both ORR and PFS had stronger associations with OS (ORR and OS: R = 0.528; 95% CI 0.081–0.798; PFS and OS: 0.778; 95% CI 0.475–0.916). Nevertheless, the upper limit of the correlation coefficient CI in both cases was higher than 0.7. Consequently, according to IQWiG recommendations, the validity of ORR and PFS as surrogate endpoints for OS is unclear [4]. In this case, the treatment effect of the surrogate in clinical trials needs to be statistically significant from the calculated STE: using 95% CI or 80% CI if the 95% CI does not lie fully above the STE. Accordingly, interventions that show a treatment benefit more than 41% ORR or 4.2 mPFS months are expected to show a significant OS benefit with sufficient certainty.

Johnson et al. [13] studied the relationship between ORR and OS in 191 trials involving patients with NSCLC receiving first- and further-line therapy. They reported a correlation coefficient of 0.40 (P < 0.0001). According to their STE calculation, a treatment benefit of 18% for 750 patients, 21% for 500 patients, and 30% for 250 patients in ORR is needed to show an OS benefit. The calculated STE incorporates observations from trials in both first- and further-line therapy. Hotta et al. [6] identified 18 phase III trials investigating EGFR-TKIs or anaplastic lymphoma kinase TKIs used as a first- or second-line treatment for NSCLC [6]. The correlation coefficient between the ORR odds ratio or PFS-HR and the OS-HR was 0.318 and 0·483, respectively. These trials, however, reported high crossover rates [24]. Similar associations between ORR or PFS and OS were observed in two other studies [10,11]. Recently, the US Food and Drug Administration published an endpoint validation study including 14 trials (12,567 patients) of first- and further-line therapy for advanced NSCLC that had been submitted between 2003 and 2013 [12]. In the trial-level analysis, there was no association between OS and ORR (R^2 = 0.09; 95% CI 0–0.33) or between OS and PFS (R^2 = 0.08; 95% CI 0–0.31). Also, in that case, included trials reported high crossover rates. In their study, no literature search was carried out to include other trials. This may translate into a likely risk of selection bias. On a broader scope, in a review, commissioned by the National Institute for Health and Clinical Excellence, of studies quantifying the relationship between PFS and OS in advanced/metastatic cancer, the relationship between PFS and OS varied considerably by cancer type and was not consistent even within one specific cancer type [25]. In summary, reported surrogacy of ORR or PFS to OS in literature is poor and in agreement with the analysis we conducted on all trials.

None of the previously published endpoint validation studies performed a stratified analysis on the basis of crossover or unbalanced postprogression treatments [6,10–13]. This is despite acknowledging that crossover and postprogression treatments may explain the observed weak associations. Thus, the reported surrogacy of ORR or PFS to OS in patients with NSCLC may be biased or underestimated in the published literature. Consequently, the reported analyses are of little or no help to decision makers when they should evaluate trials with a high treatment benefit on a surrogate endpoint. Exclusion of trials with crossover and/or unbalanced subsequent therapies appears to be a substantial factor to identify highly reliable trials. Furthermore, on the basis of the additional analysis, the existence of crossover alone seems to be the key factor that may have biased the PFS-OS relationship. The relationship between PFS and OS remained statistically significant after the adjustment for other available variables in our data set (Table 3).

This study has some limitations. We have not searched for unpublished studies. Nevertheless, we believe that we did not miss relevant studies for various reasons. First, information about crossover or postprogression treatments is mostly reported in full study reports. Thus, including, for example, an abstract database would not have added to our conclusion because all retrieved hits would have been excluded at the end. In addition, according to the funnel plots, publication bias is unlikely. Second, excluding studies from the main analysis on the basis of crossover or unbalanced postprogression treatment may have introduced selection bias. Nevertheless, the population characteristics of the trials included in different analyses appear to be comparable (Table 1), minimizing the risk of selection bias. Third, the populations, settings, and interventions included are heterogeneous. This might also affect the association between the surrogate and the true endpoints. On the basis of the multivariate model, however, the conclusion that crossover and unbalanced postprogression treatments may bias such relationship still holds. Only median age may have a minimal impact. Fourth, the definition and assessor of response is not consistent across studies and we did not account for that in our analysis. In previous endpoint literature, however, presented analyses were not stratified by response definition [6,10–13]. In addition, stratifying our analysis on the basis of response criteria would have impeded carrying out analysis in different data cuts.

Applying our methodology, that is, stratifying studies by crossover and postprogression treatments, for other surrogate endpoints in NSCLC (e.g., time to progression and duration of response), in different line settings (e.g., first-line NSCLC) or in other tumor types, is warranted. In other tumor types, it is likely that crossover and unbalanced postprogression treatments may

bias the relationship between surrogate endpoints and hard endpoints in these settings. Nevertheless, the extent of such likely bias needs to be further assessed.

Conclusions

Crossover and postprogression treatments may bias the quantitative relationship between surrogate endpoints (ORR/PFS) and OS. Therefore, the validity of ORR and PFS as surrogate endpoints for OS might be better estimated in trials that do not allow for crossover and that report balanced postprogression treatments. In second- and further-line therapy of advanced NSCLC, the validity of ORR and PFS as surrogate endpoints for OS is unclear, and a large effect size is needed to predict OS benefit with sufficient certainty. Trials that show a statistically significant treatment effect of 41% ORR or 4.2 mPFS months are expected to show a significant OS benefit with sufficient certainty. Further investigation of such methodology for other surrogate endpoints, in first-line therapy and in other tumor types and settings, is warranted.

Acknowledgment

We thank Andy Noble (Bioscript Science, Macclesfield, UK) for editorial support.

Source of financial support: This research was supported by Merck KgaA.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.07.011 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] European Medicines Agency. Guideline on the Evaluation of Anticancer Medicinal Products in Man (Report EMA/CHMP/205/95/Rev.4). London: European Medicines Agency, 2012.
[2] US Food and Drug Administration. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. 2007. Available from: http://www.fda.gov/downloads/Drugs/Guidances/ucm071590.pdf.
[3] de Sahb-Berkovitch R, Woronoff-Lemsi MC, Molimard M; participates of Round Table No. 7 Giens 2009. Assessing cancer drugs for reimbursement: methodology, relationship between effect size and medical need [Article in English, French]. Therapie 2010;65:367–72.
[4] Institute for Quality and Efficiency in Health Care. Aussagekraft von Surrogatendpunkten in der Onkologie: rapid report. 2011. Available from: https://www.iqwig.de/download/A10-05_Rapid_Report_Surrogatendpunkte_in_der_Onkologie.pdf.
[5] Soria JC, Massard C, Le Chevalier T. Should progression-free survival be the primary measure of efficacy for advanced NSCLC therapy? Ann Oncol 2010;21:2324–32.
[6] Hotta K, Kato Y, Leighl N, et al. Magnitude of the benefit of progression-free survival as a potential surrogate marker in phase 3 trials assessing targeted agents in molecularly selected patients with advanced non-small cell lung cancer: systematic review. PLoS One 2015;10:e0121211.
[7] US Food and Drug Administration. Clinical trial endpoints for the approval of non-small cell lung cancer drugs and biologics: guidance for industry. 2015. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM259421.pdf.
[8] Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228–47.
[9] Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat 2006;5:173–86.
[10] Hayashi H, Okamoto I, Taguri M, et al. Postprogression survival in patients with advanced non-small-cell lung cancer who receive second-line or third-line chemotherapy. Clin Lung Cancer 2013;14:261–6.
[11] Hotta K, Fujiwara Y, Kiura K, et al. Relationship between response and survival in more than 50,000 patients with advanced non-small cell lung cancer treated with systemic chemotherapy in 143 phase III trials. J Thorac Oncol 2007;2:402–7.
[12] Blumenthal GM, Karuri SW, Zhang H, et al. Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non-small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. J Clin Oncol 2015;33:1008–14.
[13] Johnson KR, Ringland C, Stokes BJ, et al. Response rate or time to progression as predictors of survival in trials of metastatic colorectal cancer or non-small-cell lung cancer: a meta-analysis. Lancet Oncol 2006;7:741–6.
[14] Ciani O, Buyse M, Garside R, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. J Clin Epidemiol 2015;68:833–42.
[15] Delea TE, Khuu A, Heng DY, et al. Association between treatment effects on disease progression end points and overall survival in clinical studies of patients with metastatic renal cell carcinoma. Br J Cancer 2012;107:1059–68.
[16] Flaherty KT, Hennig M, Lee SJ, et al. Surrogate endpoints for overall survival in metastatic melanoma: a meta-analysis of randomised controlled trials. Lancet Oncol 2014;15:297–304.
[17] Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
[18] Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1–12.
[19] Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions version 5.1.0. 2011. Available from: http://training.cochrane.org/handbook.
[20] Burzykowski T, Buyse M, Piccart-Gebhart MJ, et al. Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. J Clin Oncol 2008;26:1987–92.
[21] Hackshaw A, Knight A, Barrett-Lee P, Leonard R. Surrogate markers and survival in women receiving first-line combination anthracycline chemotherapy for advanced breast cancer. Br J Cancer 2005;93:1215–21.
[22] Prasad V, Kim C, Burotto M, Vandross A. The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. JAMA Intern Med 2015;175:1389–98.
[23] der Elst WV, Meyvisch P, Alonso A, Molenberghs G. Evaluation of surrogate endpoints in clinical trials (package "Surrogate" version 0.1-67). 2016.
[24] Sellmann L, Fenchel K, Dempke WC. Improved overall survival following tyrosine kinase inhibitor treatment in advanced or metastatic non-small-cell lung cancer—The Holy Grail in cancer treatment? Transl Lung Cancer Res 2015;4:223–7.
[25] Davis S, Tappenden P, Cantrell A. A Review of Studies Examining the Relationship between Progression-Free Survival and Overall Survival in Advanced or Metastatic Cancer. London: National Institute for Health and Care Excellence, 2012.